05 | 如何统一管理纷繁杂乱的数据指标?

2020-04-13 郭忆

数据中台实战课 进入课程 >

讲述：郭忆
时长 21:18 大小 19.51M

你好，我是郭忆。

上一节课，我讲到了各种类型的元数据，这些元数据有什么用？跟我们的数据中台又有什么关系呢？ 其实，元数据在指标管理、模型设计、数据质量和成本治理四个领域都发挥着作用，而这些领域构成了数据中台 OneData 数据体系。从今天开始，我将带你逐一了解元数据在上述领域的应用，首先是指标管理。

指标是一种特定类型的元数据，公司的运营会围绕它进行工作，可以说，它是业务和技术交汇点。指标数据能不能用，会影响他们的日常工作。来看一件我身边发生的事儿。

在电商业务中，新用户销售额是考核市场活动拉新效果的重要指标。马漂亮（化名）是市场部门的数据分析师，某一天，她要给 CEO 提供一份数据报告，报告中有一项指标是“新用户销售额”。孙美丽（化名）是会员中心的运营，她每天都会给 CEO 提供每日的新用户销售额数据。

结果有一天，CEO 看了这两份报告后发现，同一日的新用户销售额数值相差很大，他判断数据出了问题，责令两个部门的负责人进行排查。**排查后发现，市场部门对新用户口径的定义和会员中心不一样：**

市场部门认定新用户是首次下单并完成支付的用户；

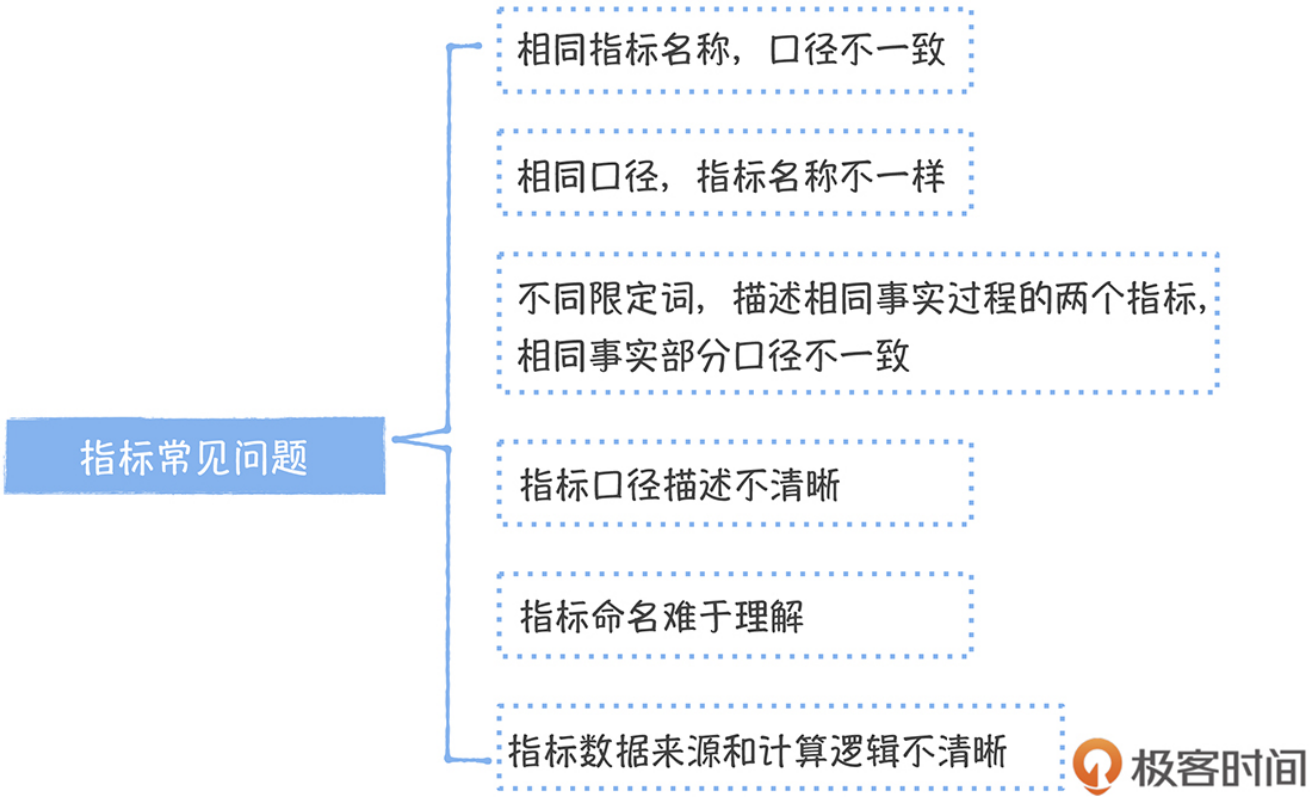
会员中心认定新用户是当日新注册用户。

也就是说，市场部门认定的新用户中，可能有之前注册但是没有下过单的客户；而会员中心只包括当日注册并完成下单支付的用户。其实，在日常工作中还有很多类似的问题。

造成上述问题的根源是因为指标口径不一致，而你要构建全局一致的指标口径，输出企业的指标字典。

指标混乱现状

从 2018 年年末开始，网易电商数据中台团队对电商业务的核心指标进行了全面的盘点和梳理，为的就是解决指标口径不一致的问题。原先 800 个指标，最终梳理完成 427 个指标，在梳理过程中，我总结了 7 个常见的指标问题，希望你能对照着看一下，自己是否也存在类似的情况。



你可以编个口诀记忆一下，比如：

同名不同径，同径不同名。

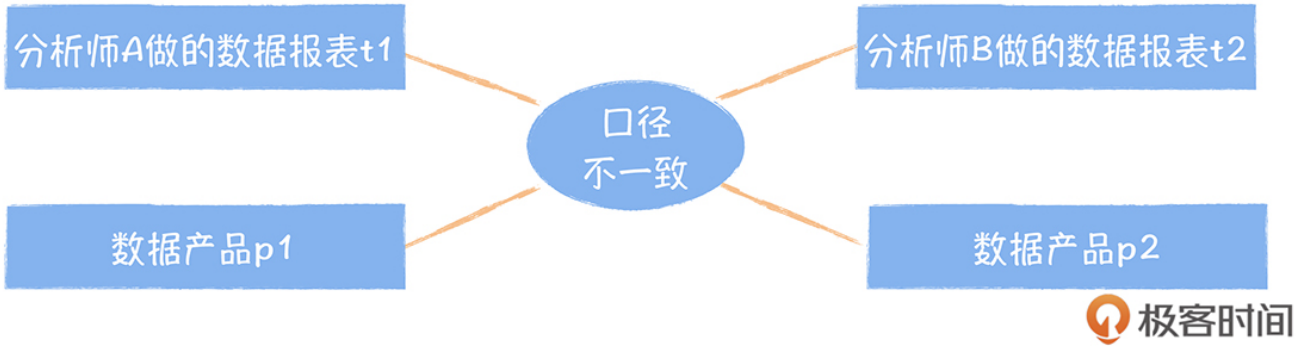
口径不清晰，口径有错误。

命名难理解，计算不易懂。

来源不清晰，同部不同径。

第一，相同指标名称，口径定义不同。

我开篇说的就是这个问题，不同的部门对相同的“新用户销售额”，因为口径定义的差别，导致指标数值的不一致。而这种情况是指标管理中最容易出现的情况。口径不一致，数据也就没办法横向对比，失去了数据辅助商业决策的意义。



第二，相同口径，指标名称不一样。

这种情况与上面相反，比如发放优惠券是电商常见的促销手段，现在你有两个数据产品：

一个是经营大脑，主要展示的是企业日常经营活动健康度的核心指标，它有一个指标叫“优惠券抵扣金额”；

一个是市场 360，主要是展示市场活动效果衡量的指标，它也有一个指标叫“优惠券消耗金额”。

其实，两者的口径定义并没有区别，但是指标名称不同，这会让使用指标的人疑惑，是不是同一个指标，计算逻辑是否一致？数据是否可以横向对比？

第三，不同限定词，描述相同事实过程的两个指标，相同事实部分口径不一致。

这个问题该如何理解呢？ 来看一个例子。

黑卡会员购买用户和非会员购买用户数，它们描述的都是用户下单购买商品的相同业务过程，记录的都是购买商品的事实，只是一个限定词是黑卡会员，一个限定词是非会员。

按照一致性原则，虽然是两个指标，但是对于购买用户数这个相同的事实部分，业务口径、计算逻辑应该是一致的，但是现实情况却可能不是这样：

“黑卡会员购买用户数”的口径定义是计算周期内去重的（重复购买的用户只算一个），下单并且支付成功的用户数量；

“非会员的购买用户数”的口径定义是计算周期内去重的，下单并且支付成功，排除关单（“关单”是指在用户在下单购买成功后，取消订单）的用户数量。

你能看到，对于购买用户数，这两个指标的口径是不一致的，一个包含关单，一个不包含关单。

第四，指标口径描述不清晰。

在梳理过程中，我们还发现，有些报表上的指标口径描述的比较笼统。比如“关单金额”，口径描述“关闭订单的金额”。不同人的理解可能不一样，有的人会认为是支付成功后关闭订单；也有可能是支付完成前，取消订单。描述不清晰，就会让人们对数据的理解产生歧义。

第五，指标口径描述错误。

在流量分析数据产品中，有“7 日 uv”这个指标，口径的定义是 7 日内日均 uv。根据口径描述的计算逻辑，应该是最近 7 日，每日 uv 相加除以 7 取平均值。显然，这个定义在业务场景中是有问题的，正确的 7 日 uv 的口径定义应该是 7 日内有登录过，去重的用户数。

第六，指标命名难于理解。

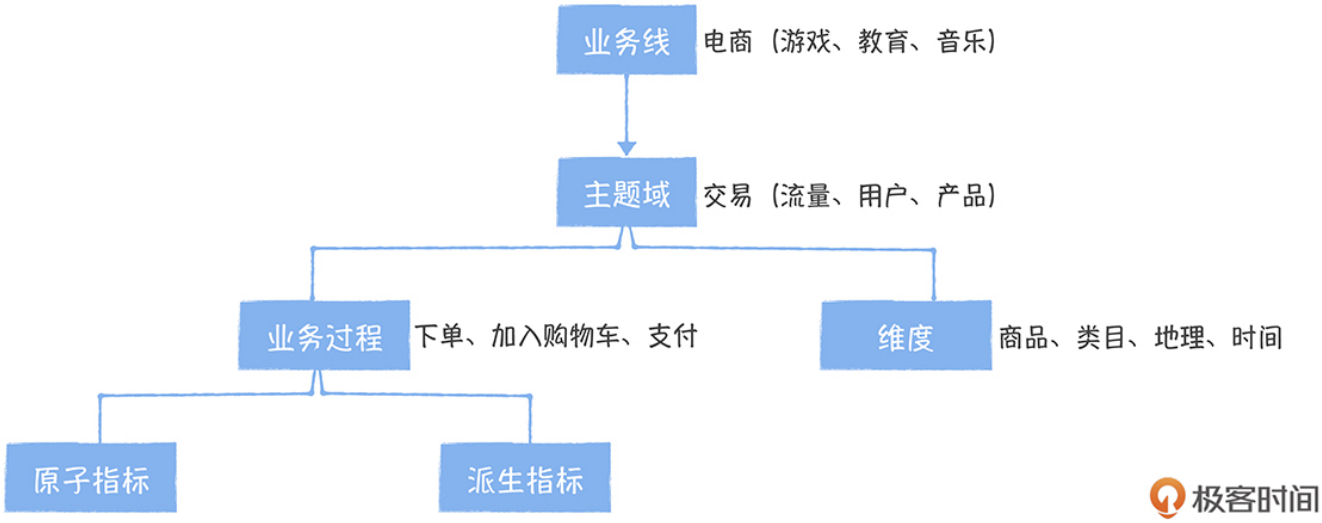
我们在梳理促销业务过程的指标时，有一个数据产品的指标名称是“ROI”，口径定义优惠券销售额 / 优惠券成本。ROI 其实是投资回报率的简称，在电商业务场景中，除了优惠券，商品降价促销都可以计算 ROI，所以比较好的命名应该是（商品|类目|通用）优惠券 ROI。所以，指标命名不规范的话，从指标名称中很难看出指标描述的业务过程。

最后，指标数据来源和计算逻辑不清晰。

如果指标数据来源不清楚，一旦这个指标数据异常，就很难去做溯源。另外，有些指标的计算逻辑比较复杂，仅仅凭借业务口径一段描述，使用指标的人还是无法理解这个指标的计算逻辑，这个时候就需要有一些伪码或者 SQL 描述。

如何规范化定义指标

那么如果你面临这些问题，该如何规范化定义指标呢？我提供给你一些经验，希望你能从中学习到如何高效、规范化的管理指标。



首先，面向主题域管理。

为了提高指标管理的效率，你需要按照业务线、主题域和业务过程三级目录方式管理指标（业务线是顶级目录）。

在网易，电商、游戏、音乐、传媒、教育都是不同的业务线。在业务线之下，是主题域，指标中的主题域与数仓中的概念是一致的，划分标准最好是跟数仓保持一致（数仓主题域的划分，我会在 06 讲详细讲述）。在主题域下面还有细分的业务过程，比如对于交易域，细分的业务过程有加入购物车、下单、支付。

其次，拆分原子指标和派生指标。

为了解决前面提到的，“黑卡购买用户数”和“非会员购买用户数”，这两个指标对购买用户数口径定义不一致的问题，我们需要引入原子指标和派生指标的管理方式。**那么什么是原子指标，什么是派生指标呢？**

统计周期、统计粒度、业务限定、原子指标，组成派生指标，所以原子指标可以定义为不能够按照上述规则进一步拆分的指标。

在例子中，你可以这样理解：

购买用户数是原子指标，原子指标的口径定义是“计算周期内去重的，下单并且支付成功的用户数量，包括关单”；

黑卡会员和非会员都可以认定为业务限定词；

统计粒度是商品粒度的；

统计周期是 30 天。

这样 30 天内，商品维度的黑卡会员购买用户数和 30 天内商品维度的非会员购买用户数就作为两个派生指标存在，但是他们继承自同一个原子指标。

除此之外，还需要指标命名规范。

指标命名规范要遵循两个基本的原则：

易懂，就是看到指标的名称，就可以基本判断这个指标归属于哪个业务过程；

统一，就是要确保派生指标和它继承的原子指标命名是一致的。

除此之外，指标应该有指标名称和指标标识（或者叫英文名）。

对于原子指标，指标名称适合用“动作 + 度量”的命名方式（比如注册用户数、购买用户数），标识的命名用英文简写或者汉语拼音缩写比较好。

对于派生指标，指标名称应该严格遵循“时间周期 + 统计粒度 + 修饰词 + 原子指标”的命名方式，标识命名要用“修饰词 _ 原子指标 _ 时间周期”的方式。

指标名称	指标标识	指标类型	主题域
购买用户数	buyer_num	原子指标	交易域
30天内黑卡会员购买用户数	black_vip_buyer_num_30days	派生指标	交易域

极客时间

第四，关联的应用和可分析维度。

对于使用指标的人（运营、分析师）了解了这个指标的口径定义之后，下一步就是要看指标的数值。所以，在全局的指标字典中，还应该有指标被哪些应用使用，这样方便去对应的数据产品或者报表上查看指标的数值。除此之外，还应该有指标的可分析维度，方便分析师从不同的维度分析指标的变化趋势。

最后一个是分等级管理。

那这么多指标，数据中台管的过来么？是的，确实管不过来，因为不仅仅是数据中台会产出一些公共核心指标，业务部门也会创建一些专属业务部门内的指标。那面对这么多指标，如何管理呢？以我的经验，你可以按照以下原则区分等级，来管理指标。

一级指标：数据中台直接产出，核心指标（提供给公司高层看的）、原子指标以及跨部门的派生指标。

二级指标：基于中台提供的原子指标，业务部门创建的派生指标。

不同等级的指标意味着管理方式不同：

一级指标，要确保指标按时、保证质量产出，指标创建由中台负责；

二级指标，允许业务方自己创建，中台不承诺指标的产出时间和质量。

现在你了解如何管理指标了吗？ 我建议你在学完这部分知识以后，结合自己所在的业务，找一些指标，试着按照上面的方法实践一下，这样掌握得会加更深刻。

指标系统

在了解如何管理指标之后，我们还需要一款好用的工具，帮助我们落实管理方法。我观察到，很多公司喜欢用 Excel 管理指标，觉得 Excel 上手容易，编辑比较方便。在我看来，Excel 并不是一个适合指标管理的工具，有这样几个原因：

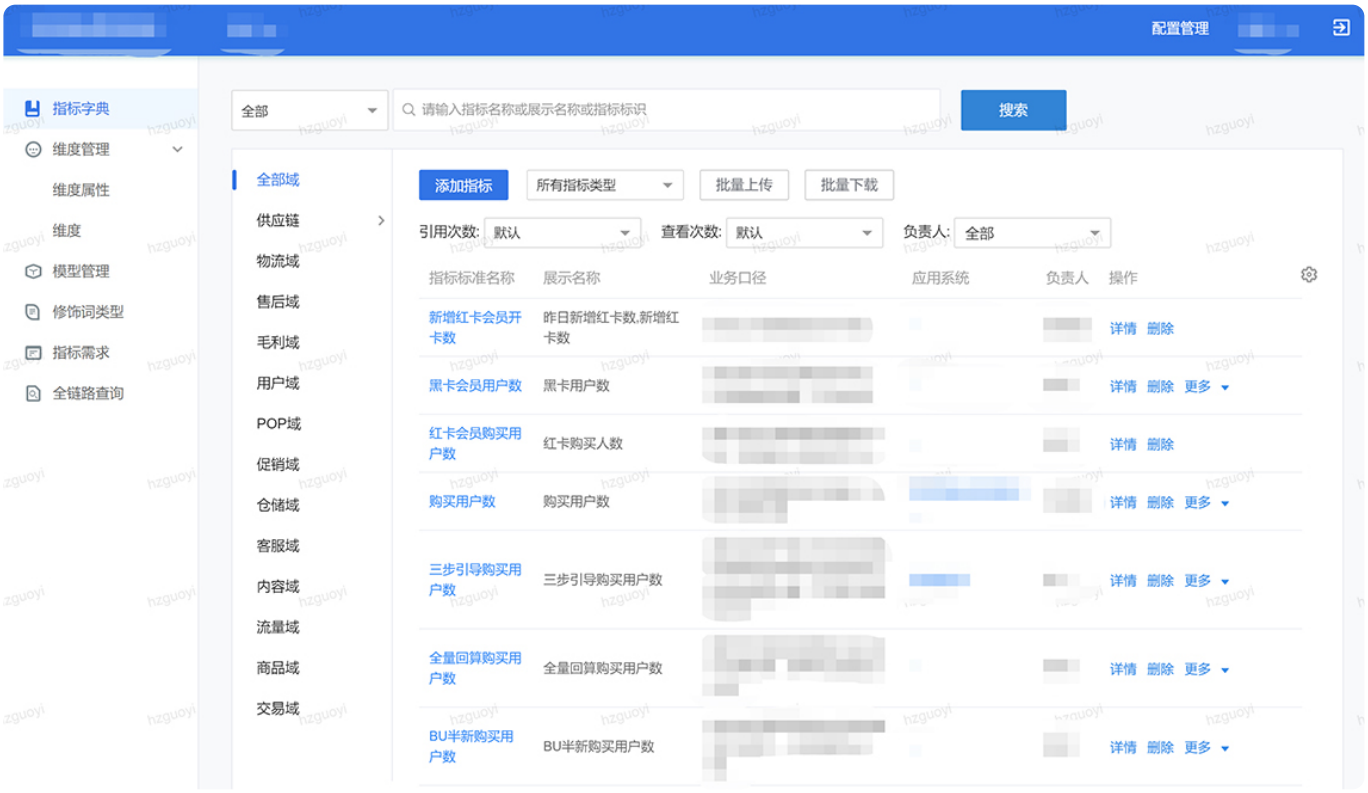
难于共享；

缺少权限控制；

无法动态更新；

指标无法跟数仓的模型动态关联。

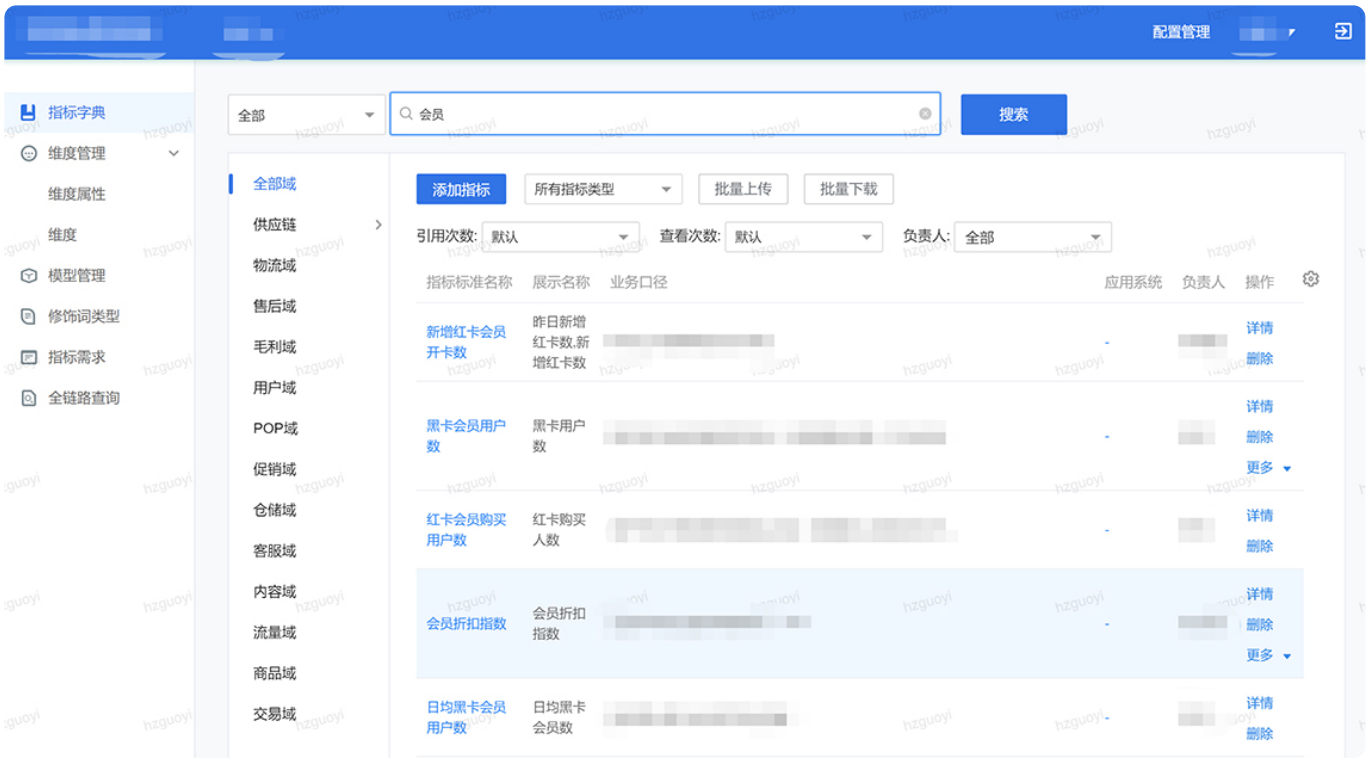
所以，我们需要一个面向指标的管理系统。



指标系统是基于元数据中心构建的一个指标管理工具，它从元数据中心自动同步数仓的主题域和业务过程，按照规范化定义创建指标。

新创建的指标同时会以特定类型的标签，下沉到元数据中心对应的表和字段上，这样在数据地图上就可以搜索到表关联的指标。

指标系统还提供了按照指标名称、标识、业务口径的检索功能。



既然指标系统能够实现指标的规范化定义，帮你解决“如何系统化、规范化定义指标”的问题，那接下来我们的重点就是如何基于指标系统构建全局的指标字典，因为这是指标治理的

最终结果。

基于指标系统构建全局的指标字典

指标治理的最终结果，就是要形成一个全局业务口径一致的指标字典。让使用指标的人，可以通过指标字典，快速了解指标的业务含义和计算过程，不会对指标口径产生歧义。

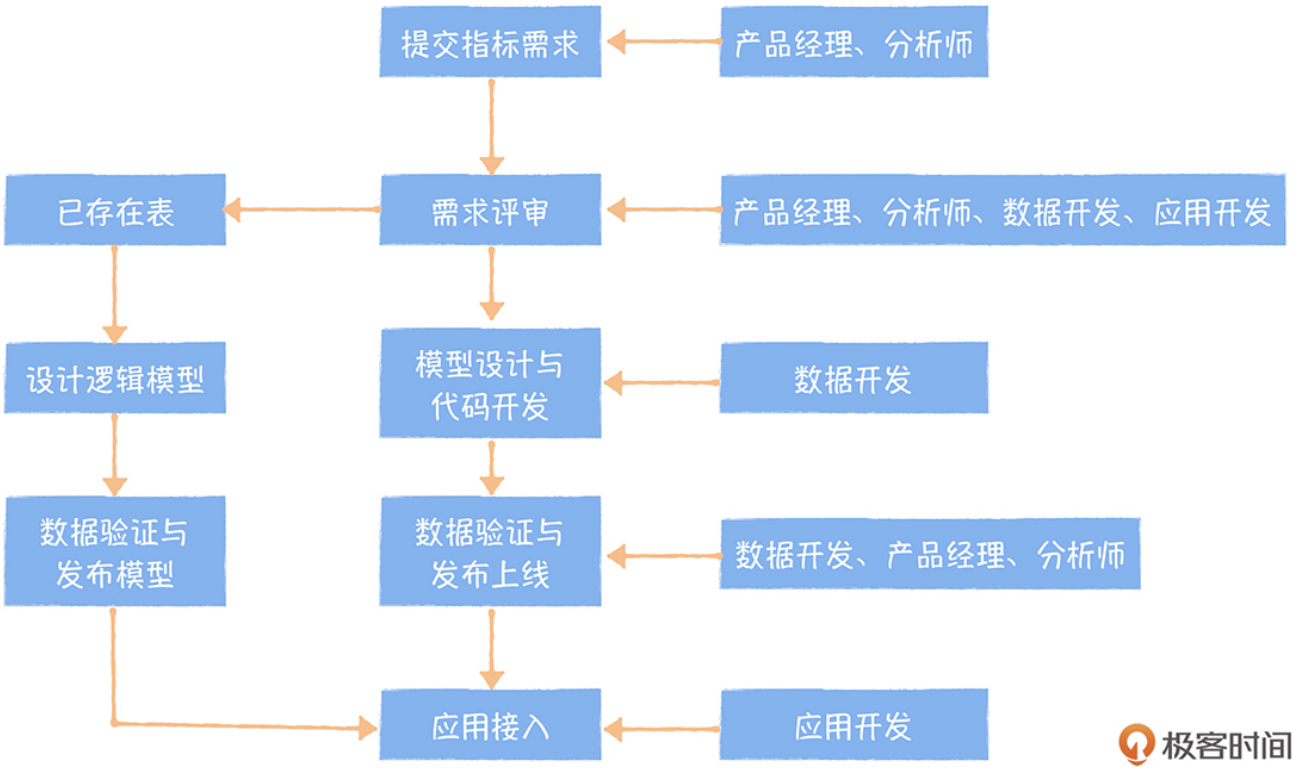
数据中台团队必须要有一个专门负责指标管理的人或者小组（一般不超过 3 个人），最好是数据产品经理来负责，**如果你的公司没有这个职位，也可以让分析师承担（前提是分析师必须属于中台团队）。**

构建全局的指标字典分为两个场景：

一个是面对一个新的指标需求，如何基于指标系统完成指标开发流程；

另外一个是面对已经存在的，混乱的指标现状，如何进行全局梳理。

先来看第一个场景。



这个图详细地描述了新建指标的流程，流程中参与的各个角色。我在这里想强调几点：

指标需求评审，需要需求方、数据开发、应用开发都参加。评审首先要确认这是不是一个新的指标，并明确它是原子指标还是派生指标。评审的目的就是要大家达成一致。

评审的结果一种是不需要开发，是一个已经存在的指标，直接可以通过设计逻辑模型（具体我会在数据服务章节讲），发布接口，获取数据。第二种就是需要开发。前者交付时间短，后者需要排期，交付时间长。

上面我提到指标有一级和二级之分，这个流程适用于一级指标，对于二级指标，可以不需要评审，当然开发也是由业务方开发和发布上线。

接下来，我们来看第二个场景。

除了新建指标的流程，对于很多公司，已经有一定的大数据业务，但是还不能算是一个中台，那这部分公司该如何进行一次全局的指标梳理呢？我认为应该有以下几个步骤：

1. 成立以数据产品或者分析师为核心的 1 ~ 3 人的工作小组，专门负责指标的全局梳理；
2. 制定指标梳理计划，明确指标梳理目标，覆盖多少个业务线，与业务方共同制定时间计划；
3. 对于每一个业务线，需要对还在使用的数据报表、数据产品进行盘点，这里顺便可以把没用的报表和数据产品应该下线；
4. 对于每一个报表和数据产品中涉及的指标，按照以下格式进行收集；

指标展示名称	购买用户数
指标标识	buyer_num
业务口径	支付成功的购买用户数量，按统计周期去重
数据来源	dwd_trd_bg_item_di
分析维度	商品、类目、流量来源、品牌
数据应用（报表）	报表连接
计算逻辑	SQL/或者伪码

极客时间

5. 对于收集的指标，明确业务口径，对于口径相同的，应该去除重复，关联的应用应该合并，此时以我的经验，可以过滤掉相当一部分；
6. 根据指标业务口径，明确指标所属的主题域、业务过程；
7. 区分指标类型，对于派生指标，要明确指标的统计粒度、修饰词、时间周期以及关联的原子指标；
8. 按照指标系统对指标的规范化定义，把整理好的指标录入指标系统。

通过全局的梳理和新建指标流程的管控，你就可以构建一个全局一致的指标字典了。

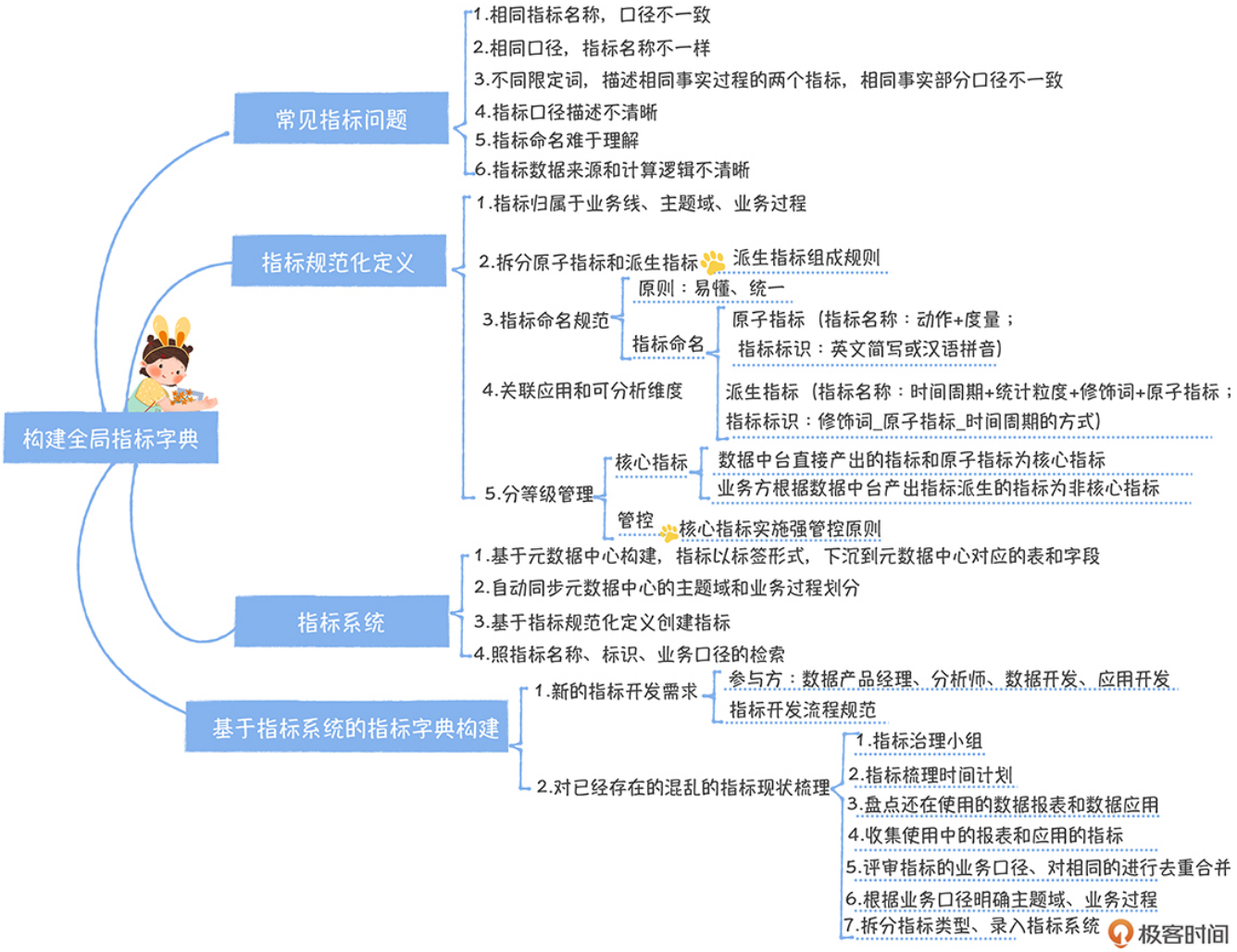
课堂总结

本节课，我带你了解了如何构建全局一致的指标字典，通过系统 + 规范的方法，帮你解决了数据中台指标一致性管理的难题，我想再强调几个点：

数据中台直接产出的核心指标必须实施强管理，由数据中台团队的专人或者小组负责，最好是数据产品经理的角色。

指标的管理必须结合系统 + 规范的治理方法，明确每个角色的职责，通过系统化的方法实现。

不同的两个指标描述的相同业务过程中的相同事实部分口径不一致，是指标梳理过程中最常见的问题，需要通过拆分原子指标和派生指标的方式解决。



思考时间

在课程的最后，还是留给你一个思考题。在一个企业的指标字典中，你觉得应该原子指标多，还是派生指标多？原因是什么呢？欢迎在留言区留言。

最后，感谢你的阅读，如果这篇文章让你有所收获，也欢迎你将它分享给更多的朋友。

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　04 | 元数据中心的关键目标和技术实现方案

精选留言 (8)

写留言

Geek_ab0a2f
2020-04-13

原子指标是根据业务流程的一些描述，而派生指标是数据需求者根据自己的需要对各种业务和业务流程的一些描述性数据，所以数仓刚建立应该是原子性指标多，随着需求的增多，派生指标就会增多

rageeeee
2020-04-13

想请问老师，有些XXX率的指标是经过计算得出的，但是感觉在业务里也不太好拆分了，像这种XXX率/XXX比例 可以作为原子指标吗?

展开

天行冕下
2020-04-13

建模理论基本上用的都是kimball的多维建模理论。
事实 --> 粒度 --> 维度 --> 指标。维度、指标都需要一致性。各个业务场景用数据仓库总线可以连接在一起。

泡泡鱼大王
2020-04-13

派生的多

展开

草裡菌
2020-04-13

自然是派生指标多，各种维度排列组合一下，就能派生出n倍。
图中还有复合指标，它的定义是什么呢？

展开

Terry郑
2020-04-13

派生指标是术，原子指标就是道。
老师这期的干活也是满满。。。看一遍基本看不懂，哈哈，还是要好好消化一下。
顺便问郭老师一个问题，数据产品经理和普通的产品经理有什么异同点吗？

展开

饭团
2020-04-13

感觉应该是派生指标多一些，因为原子指标只负责元数据，而派生指标有各种业务需求，而需求是多种多样的！所以感觉派生指标会多一些！

展开

leslie
2020-04-13

原子指标？这大概源自不可分割性吧，我们在数据系统设计时会去强调的原子性。其实老师在提出源自指标多还是派生指标多的时候可以去反向思考一下另外一个问题，我们在数据系统设计时是复合索引多还是单值索引多？
哪个多少并不重要重要的是如何合理去定义与规范？定位的规范性和合理性这才是痛点。
谢谢老师今天的分享，期待后续课程。

展开